Linguistic knowledge as a background component of an application oriented workstation

Leiter-Köhrer, Ursula

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Historical Social Research, Vol. 16 — 1991 — No. 4, 89-99

Linguistic Knowledge as a Background Component of an Application Oriented Workstation

*Ursula Leiter-Kdhrer**

Abstract: Full text systems seem often to be the cheapest way of introducing computer based methods into historical research, as, at least at first glance, the almost completely abolish the necessity for coding. It is quite frequently discovered, however, that this easy way of starting a project has to be paid for later, when the uncoded natural language makes it difficult to base results upon broad and well defined categories. Research is described which foucuses upon the introduction of formalized approaches, borrowed from linguistics. Such approaches could ultimately make the plain text, transcribed from a source, much more useful. The emphasis is put upon a concise introduction of the linguistic concepts necessary. These goals are accomplished by defining the classes of knowledge a computing environment needs to process medieval texts, as occuring in charters with a minimum of explicite coding provided.

The »historical« starting point

The starting point of our considerations are several projects at the *University of Graz,* where different editions of sources are being undertaken. (1) The textual base is unstructured running text and the information units are drawn out by hand. That is why we decided to design and implement a model for text-processing, which is able to automatically transform a text base into a factual data base. The system - to be developed - should be an integrated part of the *Historical Workstation*, introduced by Manfred Thaller (2).

A computer model, which is able to do that, needs considerable chunks of knowledge as background, including:

* Address all communications to Ursula Leiter-Köhrer, Forschungsinstitut für Historische Grundwissenschaften, Universität Graz, Heinrichstr. 26, A-8010 Graz, Österreich.

1. Linguistic knowledge
2. Historical knowledge
3. World knowledge

Only knowledge-based systems are able to make decisions, and making decisions is the only capability a computer has, because a machine would never be able to unterstand like men. (3) Therefore the system needs more formalized knowledge than human beings to process natural text.

A textprocessing system would need methods of encoding and using knowledge in ways that will produce the appropriate behavior. (4) That means to combine linguistic knowledge about all levels of description (morphology, syntax, lexicology, semantics, textlinguistics), with some important aspects of what makes humans intelligent, like knowledge about rules, objects, actions, events and heuristics, and knowledge about the (historical) reality.

We propose that a nonstructured text could be structured by using several domains of knowledge, particularly knowledge in linguistics. This way of analysing we call »semantic parsing» (5). The concept of »semantic parsing« is based on the theory of reproduction of (historical) reality in sources and the theory of symbols and their use in greater structures. To parse a text according to a semantic concept means the tokenization of running text into units, which are not only defined by their form, but also - and that is very important - by their meaning. (6) Semantic knowledge is not to be considered stand-alone, but always in connection with other aspects of signs (syntax) and the situation in which the sign is used (pragmatics).

The semiotic model

Each communication is based on using signs and the text is the verbal output of a communication-event. But a text is not only the reproduction of an (historical) event or an intention of an author, the text itself is also an actual reproduction of an abstract semiotic model. Therefore it's obviously clear, that the phenomenon »sign« is the focal point of our reflections.

A sign is »something«, that has a perceptible appearance and is vicarious for something else. Depending on the kind of perception and the kind of relationship to the object or fact of the real world, different classes of signs are distinguished. A language sign is a non-natural (7) visual or acoustic symbol (8) and its formal representation is arbitrary connected to the meaning of the sign. There is no reason why the symbol *five* has the meaning of *a size between four and six*; that is only a convention within a group of people.

Concerning signs there are some important aspects, which should be taken into consideration for the development of a semiotic model suitable for implementation:

- Once Charles Morris (9) defined three aspects, which are still fundamental for an adequate description of signs and the interrelations between them. This model is called a »semiotic triangle«.
    1. The syntactic aspect: A sign is always connected to other signs. This relation organizes on one hand the relationship between signs of the same level, the same complexity, and on the other hand the principles of generating signs more complex in structure (sign - sign relation).
    2. The semantic aspect: A sign is always connected to an object, a fact or an event of the real world - or the miniworld -, which should be reproduced (sign - object relation).
    3. The pragmatic aspect: A sign is always connected to its user. This relation brings up the contextual components using signs, the intention of the author, the historical background and the immediate situation (sign - user relation).

- According to the meaning of structural linguistics the language is considered to be a system of signs. The next step is to describe the position, which a sign may have within the system.
    1. Between signs there is a syntagmatic relationship, which organizes the combination of signs in their linear sequence. Signs which are able to be combined to generate more complex structures - like sentences - are called to be in *contrast.* Not every sign is able to arise at every position of a sentence.
    2. There is a paradigmatic relationship between signs of the same class. Elements of one class are in *opposition,* hence they are not allowed to be combined for producing greater structures. Only one element can appear at a special position within the sentence, but each member is able to replace an element of the same class. The membership of a generalized class is an inherent property of signs.

A	tall	man	opens		the	black	door
The		woman	closes		a		window
		She	walks	up	the		street

⇕ paradigmatic relationship
⇔ syntagmatic relationship

- The sign itself consists of two inseparable parts: the formal representation, the perceptible appearance and the meaning, the sense given to the sign. This dualism is the starting point of many problems in describing signs, because the relationship between these two parts is dynamic and not unambiguous. Depending on the unambiguity and the dynamics of the »form - meaning ratio« there are different types of relations:

1. The 1:1 relation means that only one formal representation is connected to exactly one proposition.

   form } meaning

   Regarding the dynamics of the lexicon, depending on time and place, this kind of relation is quite rare.

2. The n:1 relation means that different formal representations have the same proposition, they are the concrete variants of one abstract phenomenon. Particularly in historical texts a great richness of variants is expected, especially the names of persons appear in different spellings. (10)

   form_1
form_2 } meaning
form_3

   Before processing all these variants should be equalized by suitable methods, like the *soundex* -algorithm (11), the *skeleton* -algorithm (12) and the algorithm for conversions (13), which have already been implemented within the Historical Workstation. (14)

3. The 1:m relation means that one formal representation can have different propositions.

   form { meaning_1
meaning_2
meaning_3

   This is one of the typical features of natural language: the ambiguity of language signs. The disambiguation is possible by charac-

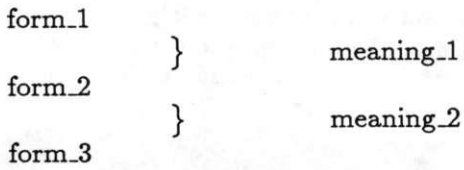
teristics of the context and co-text. Different kinds of ambiguity are possible:

(1) the grammatical ambiguity (e.g. the Latin form *exercitus* has six different meanings),

(2) the syntactic ambiguity (e.g. the sentence *John told Robert's son that he must help him.* may be read in six different ways.) and

(3) the lexical ambiguity (e.g. within the phrase *Ulricus de Kirchbach* the word *de* indicates either *Ulricus comes from a place named »Kirchbach«* or *Ulricus is a member of the gentry* or *Ulricus is a member of a family named »de Kirchbach«* or different combinations of the aforementioned possibilities).

4. The n:m relation means that different formal representations and different meanings are overlapping.



- The meaning of a sign itself can be described as a cluster of semantic features (semantic markers). Those are the smallest parts into which the meaning of a sign could be divided (e.g. U+ /- humane, U+ /- livinge, U+ /femalee). Semantic markers are used for generalizing concepts and for building semantic trees and networks. The nodes represent the concepts including the semantic features and the edges represent the possible relations between concepts and the hierarchical dependency. Semantic features are responsible for semantic restrictions, that means that some concepts are not to be combined with some others. For instance a concept with the feature U+ livinge cannot produce a greater structure with a concept containing the feature U- livinge. (15) In case of lexical ambiguity the semantic markers make possible the disambiguation. If the word *bachelor* occurs combined with an adjective containing the marker U+ femalee (e.g. *pretty),* you can be sure that the meaning of *bachelor* is not *a man with no woman,* but well *a female person with an academical degree.*

An adequate description of signs (symbols) of different levels of complexity is only possible if all these features are connected to a semiotic model of description. A text is the most complex structure of signs bound together.

Historical Social Research, Vol. 16 — 1991 — No. 4, 89-99

The Text - a cluster of relations

Each current text is an actual variant of an abstract text-class-concept, which can be described in its prototypical macro-structure. Depending on this structure several kinds of entities and relations are possible and expectable.

Entities are continuous or discontinuous units which consist of one or more signs with different levels of complexity. These conceptional entities can be described by their semantic and syntactic properties as well as by their semantic and syntactic relations to other entities.

The relations ensue from

(1) linearity of text,
(2) the syntactic and semantic dependency and
(3) the paradigmatic substitution-frame.

The current text may now be seen as a cluster of relations. The nodes of the network are the actual entities and the edges are the different actual relations between them. The entities and the relations within the network can be seen as the formal reproduction of the current linear text. This might be seen as a »hyper-representation» (16) of text.

The levels of description

For the implementation of such a semiotic model as background component in an application oriented workstation, the system ought to provide several levels of description. This process is controlled by the user in a twofold process: On the one hand there is the formal semiotic description of the entities to be found and on the other hand there is a set of rules, which binds these descriptions onto patterns, which can be parsed in the running text.

All definitions are made by the user. Only some sets of characters are predefined, like the standard alphabet, the standard digits and the standard terminating signs, nevertheless it's possible to change these pre-assumptions.

The »semantic parser« has two separate parts: (17) (1.) a declarative part for defining the entities and their properties and (2.) a procedural part containing the production rules. The first part is principal data-independent, because the descriptions represent generalized concepts. The second one is more data-dependent, because the rules bind the descriptions onto the current text. This unification results in the individual entities of the text, which could be (1.) marked with user-defined start/stop symbols, (2.) simply listed or (3.) put into an external dictionary.

The **declarative component** of the »semantic parser« provides several levels of description. For the identification of entities the user has to declare them with all their relevant properties. They can be static, if they are inherent features and hence context-independent, and they can be dynamic and fuzzy, if they related to the situation (time and place), to other entities found or to the characteristics of the context or co-text. Within these declarations several tools are supplied, like external dictionaries, semantic networks and different methods, for instance the algorithms for the equalization of variants described above.

The levels of description depend on the complexity of the entity to be indicated. First the single signs have to be declared, afterwards the more complex signs, which use specified sets of single signs, and at last the units (entities), which are to be found in running text. To make this possible, the »parser« supplies several commands, each command representing one level of description. The commands support either the description of the perceptible appearance, or the formal description of the meaning, or both of them.

level	command	form	meaning
0	set	+	–
1	pattern	+	–
2	type	+	+
3	entity	+	+
4	context	+	+

SET: *»Sets«* are set-theoretical constructions used to specify characters (single signs) as elements of a *set* or a *subset.*

There are some predefined sets, like the standard characters of an alphabet (named *alpha),* the standard delimiting signs (named *delimiter)* and the standard digits (named *digit).* Other sets, like the set *vowel,* is to be defined by the user.

PATTERN: *»Patterns«* are formal descriptions of units within the running text.

They are constructed either with constants, or with defined *sets.* There are predefined *patterns,* like the standard assumptions in case of the units *word* and *sentence,* they are defined by the predefined *sets.* The *pattern*-command supports several pattern-matching routines. (18) *Patterns* only act on the level of formal description, there is no aspect of meaning involved.

TYPE: *»Types«* are also set-theoretical constructions, but unlike *sets,* the members are objects or patterns (objects defined by their formal representation).

A *type* can be seen as a conceptual user-defined class and the members are

the possible extension of such a class. Therefore the membership brings up a semantical component of description. *Types* are either defined by enumerating all possible members, or - if the number is too great - by defining the *type* as a pointer to a node of an external semantic network. The extension of a *type* defined like this are all objects, which are related to the node by the *is-a* relation.

ENTITY: *»Entities«* are the most complex units, which are possible to be declared.
The definition of such *entities* encloses the description of all properties a entity may have. Unlike *types* which are described in an extensional way by enumerating, the *entities* are described intensionally by defining their features. Like *types* an *entity* may refer to an external dictionary or a semantic network and by this way an extensional component of description is involved. An *enitity* is described by a typical combination of different features, like:

- the syntactic features, which specify the syntactical behaviour (syntactical function) and the syntactical rules for combination,
- the semantic markers, which formalize parts of the enitity>s meaning. Sometimes it is very difficult to distinguish between syntactic and semantic features (19),
- the semantic relations, which either relate entities or collate them to conceptual classes,
- the paradigmatic properties, which mean the membership of a conceptual class or a pointer to an external dictionary or network,
- the syntagmatic properties,which define the position within the linear sequence (text), the number and the position of smaller units generating the entity,
- the levels of complexity by declaring all smaller units, which are components of an entity, and the dependency between them. Such units are e.g. *types, patterns* or as well *entities,*
- the formal appearance, defined by the *pattern* -statement.

The actually found units in the current text represent the actual extension of the declared entity.

CONTEXT: The *context* -statement is used twice: For the makro-structure on the one hand (*co-text)* (20), and on the other hand for the contextual implications (*context),* if the text-production and text-comprehension is influenced.

The declarative component itself is insufficient for indicating user-defined units in current texts. The matching routines act with a production system, the **procedural component** of the »parser«.

Even this component is defined by the user, who declares sets of rules which bind the description onto patterns, possible to be parsed in the text. A single rule consists of two parts: the condition and the conclusion (21) A condition »fires«, if the pattern matches. If one user-defined unit is found and the unit is a member of a greater structure, the system tries to »complete« the unit. In case of success the unit will be indicated, in case of failure the system tries another rule. This step by step process will be done until there is no rule possible anymore.

Notes

(1) (1) The *Urkundenbuch der Steiermark und ihrer Regenten.* See the recent report on the project: Friedrich Hausmann: Urkundenbuch der Steiermark und ihrer Regenten, Band I - III und V ff. In: XII. Bericht der Historischen Landeskommission für Steiermark, 16. Geschäftsperiode (1982 - 1986), ed. by Othmar Pickl. Graz 1988, pp.79 - 90. (2) The *urkundenbuch des Patriarchats Aquileia.* The first volume of this project is already published: Reinhard Härtel: Die ältesten Urkunden des Klosters Moggio (bis 1250). Wien 1986 (Publikationen des Historischen Instituts beim Österreichischen Kulturinstitut in Rom, 2. Abt., Reihe 6), (3) The Research on the medieval administration and chancery of Regensburg: See Susanne Botzem - Ingo H. Kropa£: Integrierte Maschinelle Edition am historischen Arbeitsplatzrechner: Repräsentation und Dokumentation von Quellen und Wissen am Beispiel des Regensburger Kanzlei- und Verwaltungswesen im Spätmittelalter. Graz, Institut für Geschichte 1989 (Integrierte Maschinelle Edition - Bericht 1, unveröff. Arbeitspapier).

(2) See Manfred Thaller: The Daily Life in the Middle Ages, Editions of Sources and Data Processing. In: Medium Aevum Quotidianum Newsletter 10, 1987, pp. 6 - 28.

(3) See the discussion between the representatives of the *»Strong AI assumption«*, like Douglas R. Hofstadter (ed.): The mind's I. Fantasies and Reflections on Self and Soul. Toronto 1981, who says that »Minds exists in brains and may come to exist in programmed Machines« (p.382.), and the representatives of the *»Wake AI assumption«*, like (1) John Searle: Minds, Brains, and Programs. In: Hofstadter (ed.) 1981, (2) Terry Winograd and Fernando Flores: Understanding Computers and Cognition. A New Foundation for Design. New Jersey 1986.

(4) See James Allen: Natural Language Understanding. California 1987.

(5) See Ingo H. Kropad - Ursula Leiter-Köhren Analytical Semantic Parsing System: Ein Programm zur automatisierten Indizierung und In-

haltserschließung. Graz, Forschungsinstitut für Historische Grundwissenschaften 1989 (ASPS - Bericht 1, unveröff. Arbeitspapier).

(6) Originally the term »parsing« was used for processing sentences only into their syntactic structure, during the last years a semantic interpretation may be involved.

(7) Unlike non-natural signs, the formal representation of a natural sign is given by nature and not depending on convention, e.g. *smoke means fire.*

(8) According to the »form-meaning relationship« there may be a formal coherence *(icons),* a causal coherence *(indices)* or a conventional coherence *(symbols).*

(9) See Charles W. Morris: Foundation of the Theory of Signs. Chicago 1938.

(10) In the *Prosopographical Databank for the History of the South-East Territories at the Roman Empire until 1250* there are 1985 persons mentioned, whose name has the meaning »Ulrich«, but this single meaning appears in exactly 91 different orthographical variants.

(11) The *soundex* -algorithm transforms a source-string into a numeric value, depending on the phonetical value of each character of the string.

(12) The *skeleton* -algorithm reduces a source-string into a skelet of characters, which are most distinctive in sense.

(13) The *conversion-tool* consists of several sets of position-sensitive rewriting-rules. The source-string is systematically and gradually converted into a »normalized« target-string.

(14) See Manfred Thaller: Kteico 3.1.1. Ein Datenbanksystem. St. Katharinen 1989 (Halbgraue Reihe zur Historischen Fachinformatik, Serie B: Softwarebeschreibungen, Bd. 5); Peter Becker: KXEICO. Ein Tutorial. St. Katharinen 1989 (Halbgraue Reihe zur Historischen Fachinformatik, Serie A: Historische Quellenkunden, Bd. 1).

(15) At this place the well known sentence of Noam Chomsky should be mentioned: *Colourless green ideas sleep furiously.*

(16) See Josef Wallmannsberger: Hypertextmodelle in der Informationswissenschaft: Die Welt als Text - Die Bibliothek als Hypertext. In: Mitteilungen der Vereinigung Österreichischer Bibliothekare 43, 1990), pp. 6 - 19, decribes the four distinctive features of hypertext: (1) Associative chains, (2) different data-types (visual and acoustic) should be an integrated part of the textbase, (3) more possibilities in information retrievel and a (4) pragmatic view of actions and situations

(17) The following explications represent the actual state of the research in the project titled *»Systementwicklung zur automatisierten Indizierung und Inhaltserschließung für Zwecke der historischen Faktendoku-*

mentation (Analytical Semantic Parsing System - ASPS)«, sponsored by the Austrian *Ministerium für Wissenschaft und Forschung.*

(18) See Wolfgang Levermann: CMATCH: Mustererkennung in Zeichenketten. St. Katharinen 1989 (Halbgraue Reihe zur Historischen Fachinformatik, Serie B: Softwarebeschreibungen, Bd. 4).

(19) See Noam Chomsky: Aspekte der Syntax-Theorie. Frankfurt am Main 1978.

(20) For the particular formulas and the prototypical semantic structure of charters see: Ingo H. Kropa<5: Informationssysteme in der Geschichtswissenschaft. Konzeption und Anwendung am Beispiel der Prosopographischen Datenbank zur Geschichte der südöstlichen Reichsgebiete bis 1250 (PDB). Graz 1988 (dzt. unveröff. Habilschrift)

(21) See Cosima Schmauch: Wissensrepräsentation. In: Thomas Christaller (ed.): Künstliche Intelligenz. 5. Frühjahrsschule, KIFS-87, Proceedings, Berlin 1987.